Computer Architecture
Lecture 1
Fundamental Concepts

What is Computer Architecture?

- The science and art of designing, selecting, and interconnecting hardware components and designing the hardware/software interface to create a computing system that meets functional, performance, energy consumption, cost, and other specific goals.

Why Study Computer Architecture?

Why Study Computer Architecture?

- Enable better systems: make computers faster, cheaper, smaller, more reliable, …
  - By exploiting advances and changes in underlying technology/circuits

- Enable new applications
  - Life-like 3D visualization 20 years ago?
  - Virtual reality?

- Enable better solutions to problems
  - Software innovation is built into trends and changes in computer architecture
    - > 50% performance improvement per year has enabled this innovation

- Understand why computers work the way they do

Computer Architecture Today

- Today is a very exciting time to study computer architecture

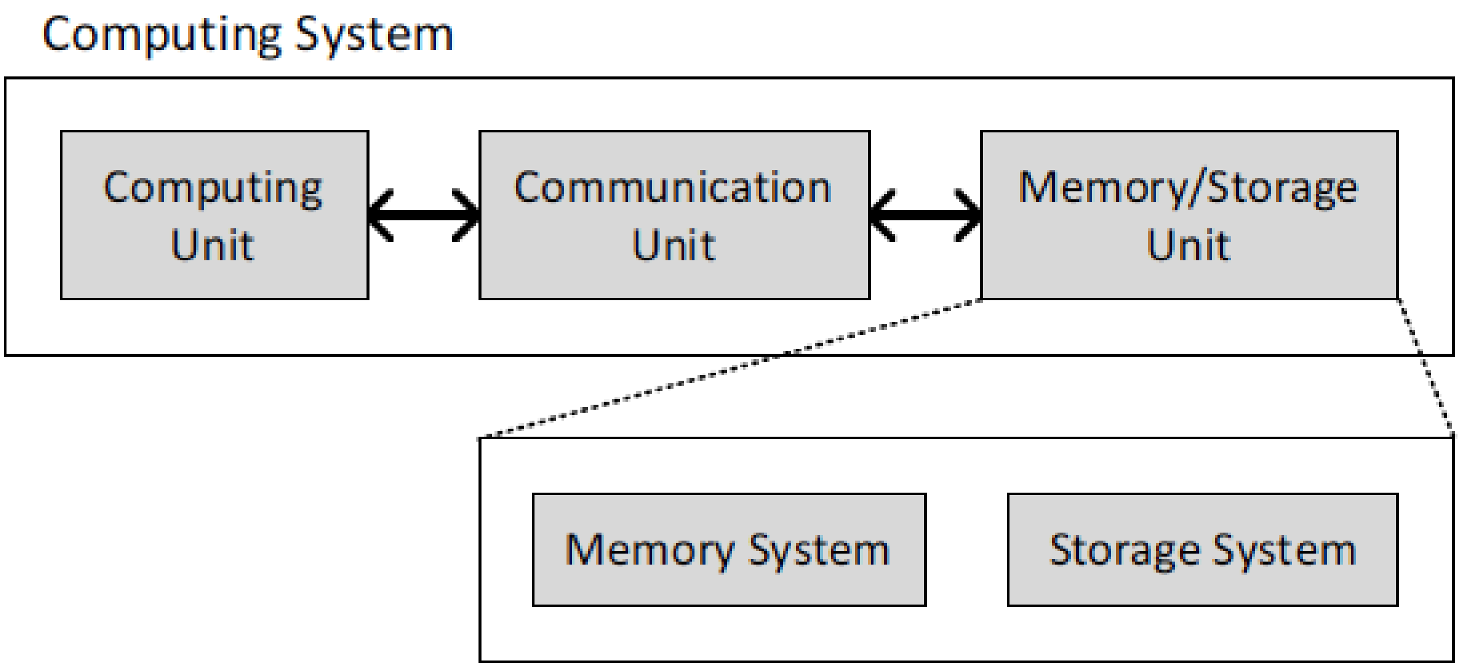
- Industry is in a large paradigm shift (to multi-core and beyond) – many different potential system designs possible

- Many difficult problems *motivating* and *caused by* the shift
  - Power/energy constraints
  - Complexity of design → multi-core?
  - Difficulties in technology scaling → new technologies?
  - Memory
  - Reliability
  - Programmability

What is A Computer?

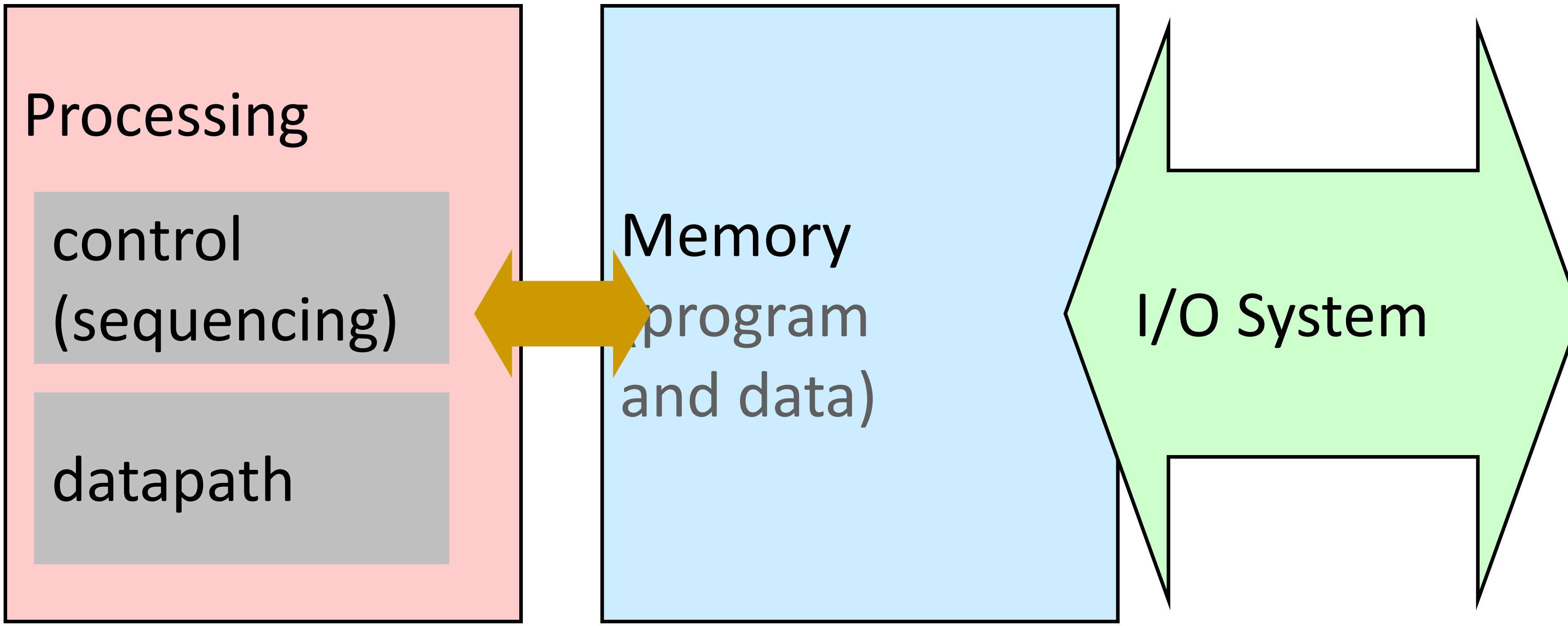
Three key components in computer system

- Computation
- Communication
- Storage (memory)

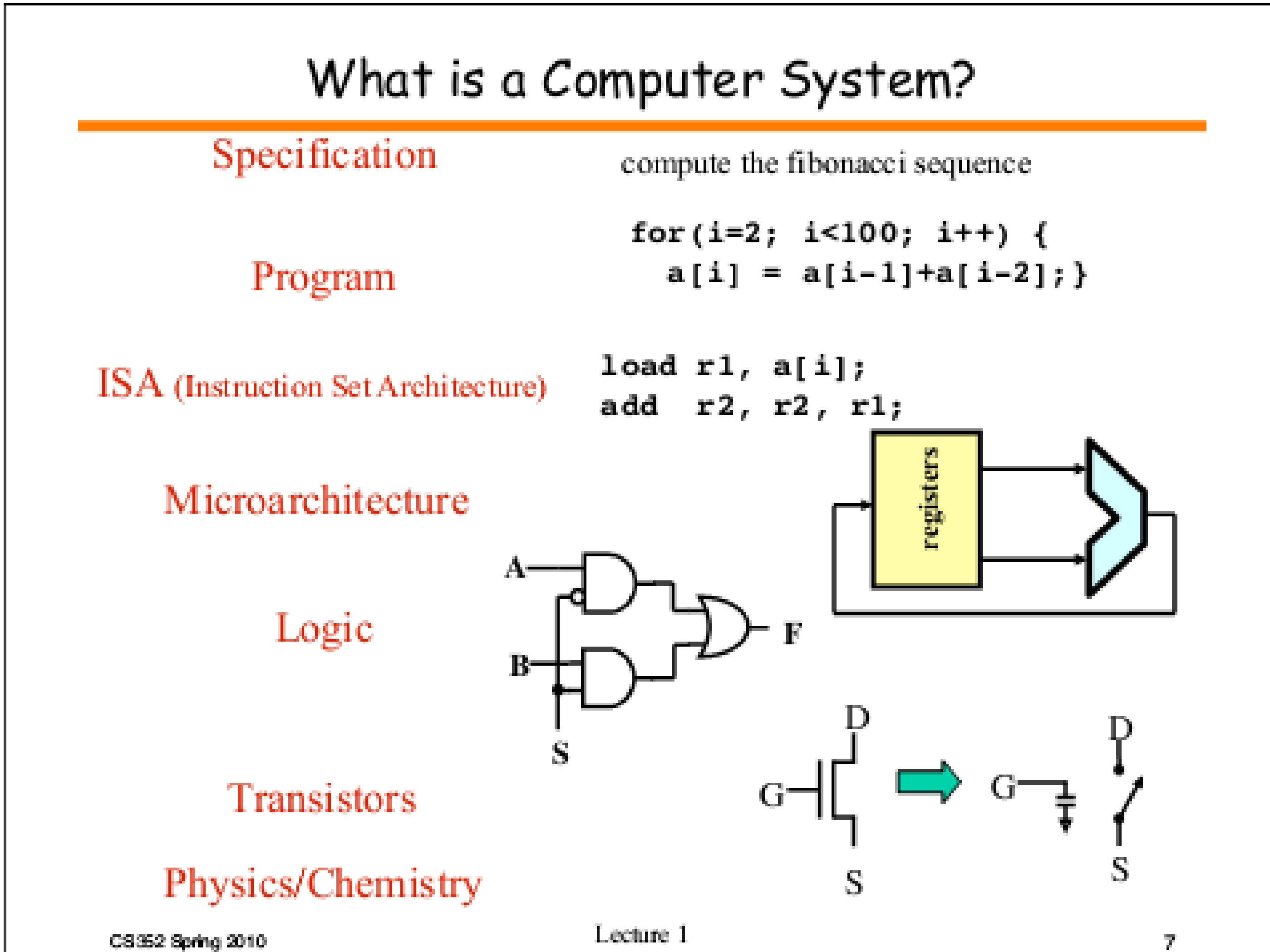


What is A Computer?

- We will cover all three components



What is a Computer System?



Levels of Representation

High Level Language Program

```
temp = v[k];
v[k] = v[k+1];
v[k+1] = temp;
```

Compiler

Assembly Language Program

```
lw   $15, 0($2)
lw   $16, 4($2)
sw   $16, 0($2)
sw   $15, 4($2)
```

Assembler

Machine Language Program

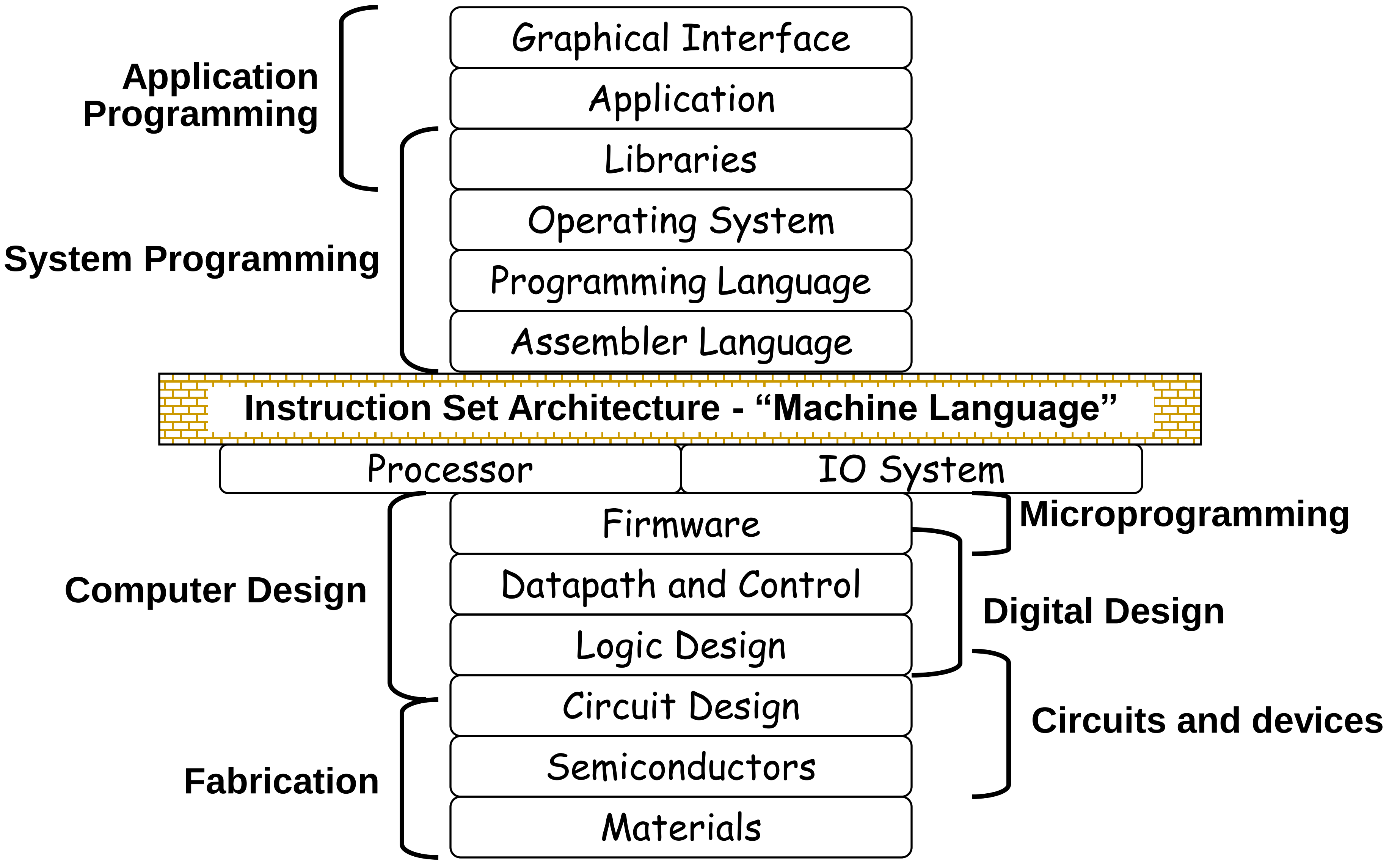
```
0000 1001 1100 0110 1010 1111 0101 1000
1010 1111 0101 1000 0000 1001 1100 0110
1100 0110 1010 1111 0101 1000 0000 1001
0101 1000 0000 1001 1100 0110 1010 1111
```

Machine Interpretation

Control Signal Spec

```
ALUOP[0:3] <= InstReg[9:11] & MASK
```

Levels of Abstraction



Computer Organization

- Capabilities & Performance Characteristics of Principal Functional Units

(e.g., Registers, ALU, Shifters, Memory Management, etc.

- Ways in which these components are interconnected
  - Datapath - nature of information flows and connection of functional units
  - Control - logic and means by which such information flow is controlled

"Hardware" designer's view includes logic and firmware

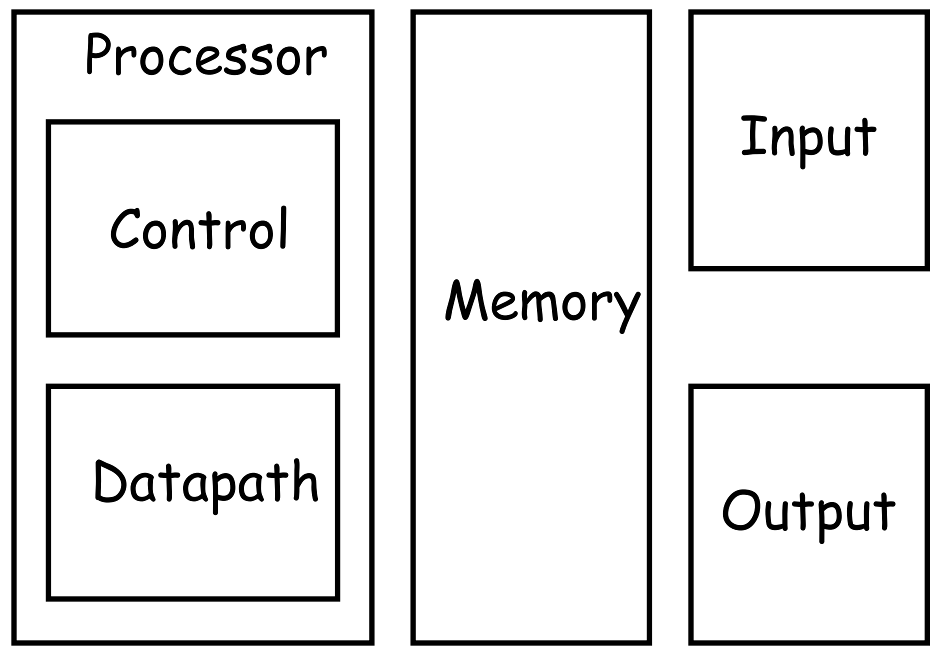
This Course Focuses on General Purpose Processors

- A general-purpose computer system
  - Uses a programmable processor
  - Can run "any" application
  - Potentially optimized for some class of applications

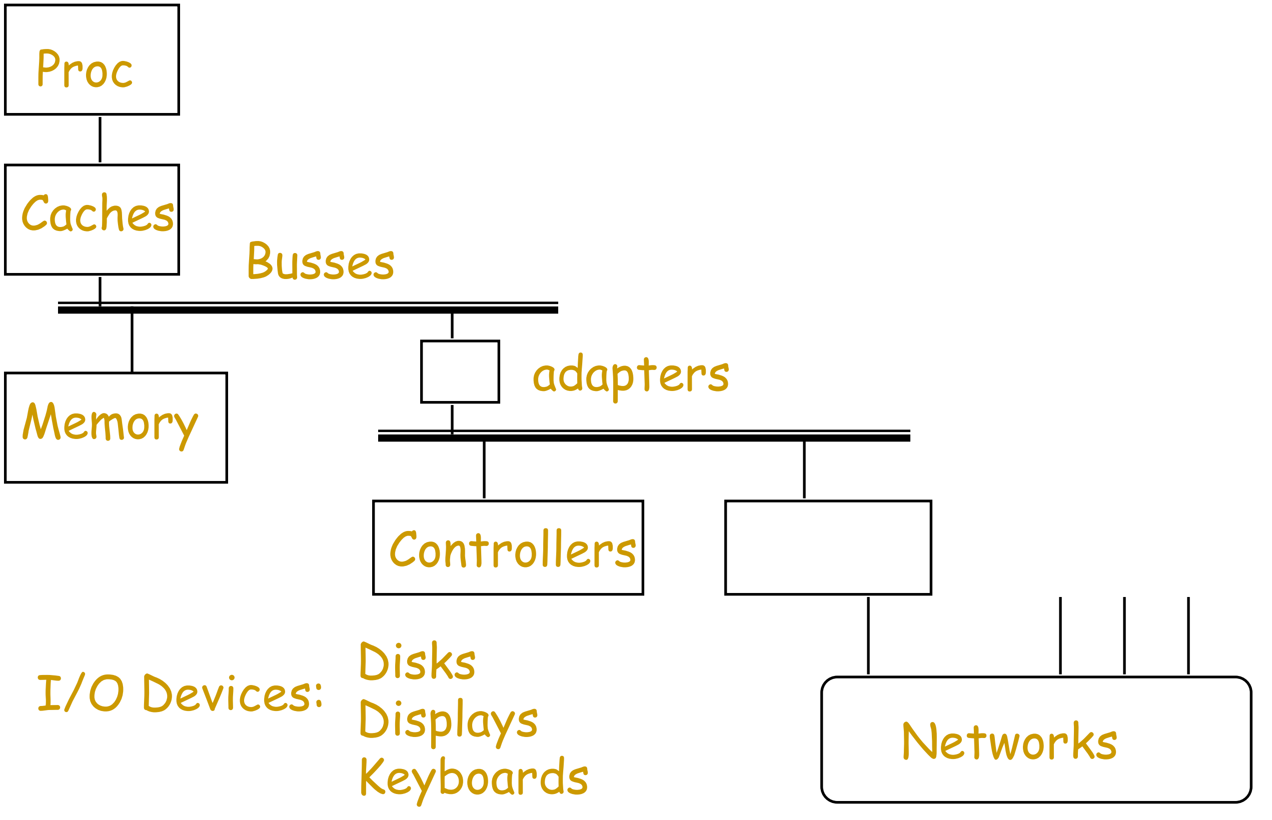
Unified main memory

- For both programs & data

Busses & controllers to connect processor, memory, IO devices

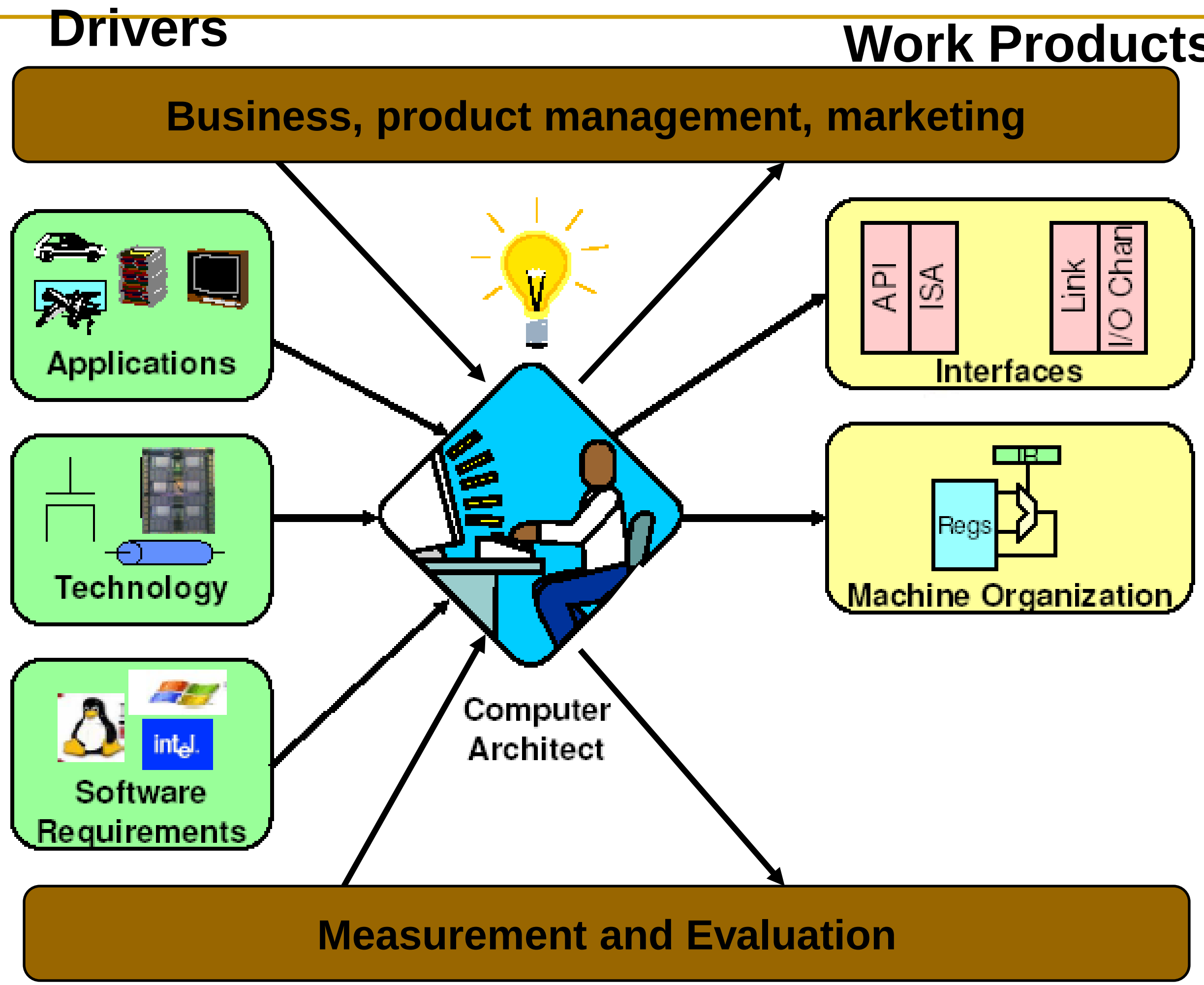


Today, "Computers" are Connected Processors



- All have interfaces & organizations

What does a computer architect do?



Translates business and technology drives into efficient systems for computing tasks.

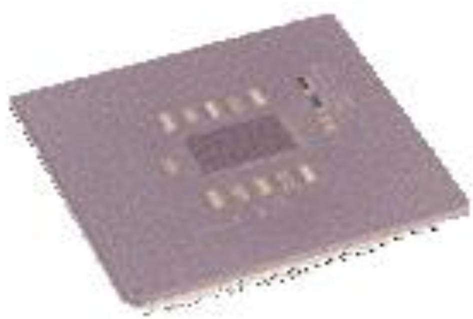
What is A Computer?

- **<u>Input/Output (I/O) devices</u>**– These allow you to send information to the computer or get information from the computer.



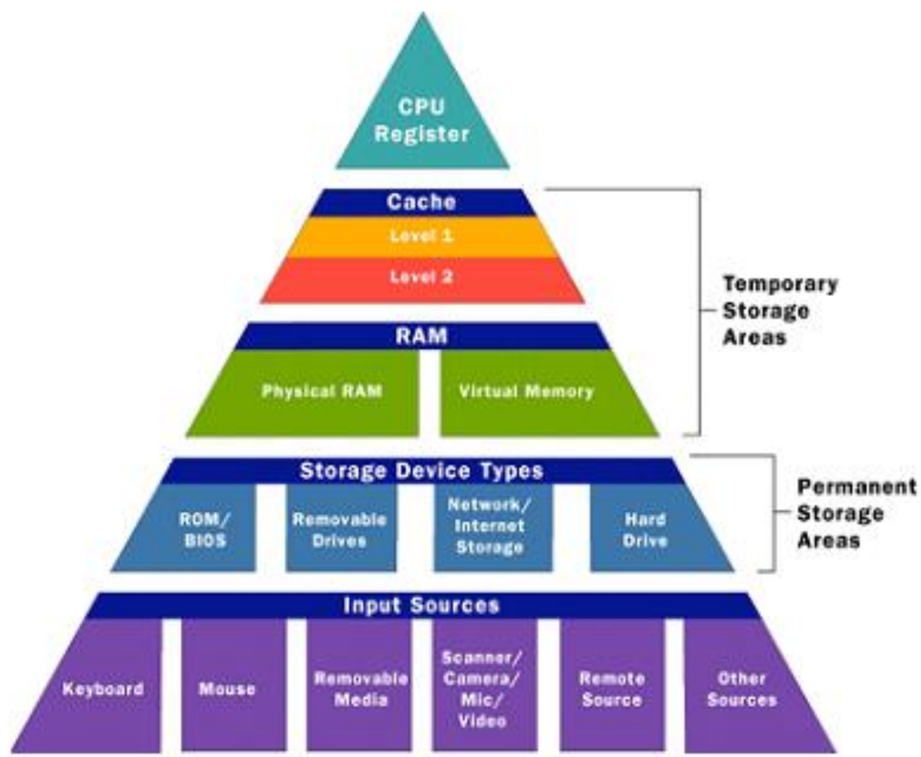
What is A Computer?

- **Central Processing Unit–** CPU or Processor for short. The brain of a computer. Approximately 1.5 in X 1.5 in. Does all the computation/work for the computer.



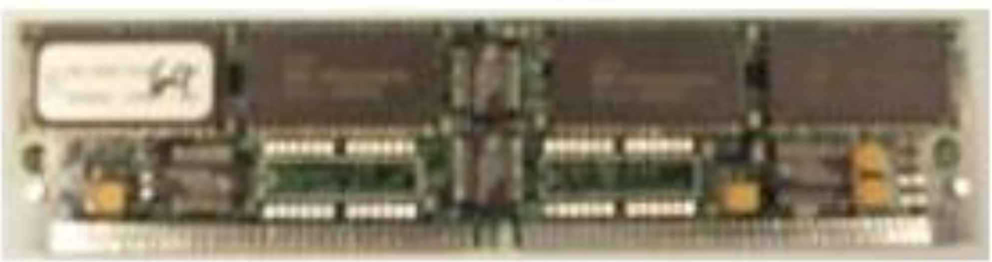
What is A Computer?

- **<u>Memory– </u>**Although memory is technically any form of electronic storage, it is used most often to identify fast, temporary forms of storage. Accessing the hard drive for information takes time. When the information is kept in memory, the CPU can access it much more quickly.



What is A Computer?

- **Random Access Memory– RAM.** Where information is stored temporarily when a program is run. Information is automatically pulled into memory, we cannot control this. RAM is cleared automatically when the computer is shutdown or rebooted. RAM is volatile (non-permanent**).**



What is A Computer?

- **Read Only Memory– ROM.** More permanent than RAM. Data stored in these chips is nonvolatile -- it is not lost when power is removed. Data stored in these chips is either unchangeable or requires a special operation to change. The BIOS is stored in the CMOS, read-only memory.

What is A Computer?

- <u>**Hard Drive–**</u> Where you store information permanently most frequently. This is also nonvolatile.

What is A Computer?

- **<u>Motherboard</u>** – A circuit board that allows the CPU to interact with other parts of the computer.

What is A Computer?

<u>Ports –</u> Means of connecting peripheral devices to your computer.

- **Serial Port** – Often used to connect a older mice, older external modems, older digital cameras, etc to the computer. The serial port has been replaced by USB in most cases. 9-pin connector. Small and short, often gray in color. Transmits data at 19 Kb/s.
- **Monitor Ports** – Used to connect a monitor to the computer. PCs usually use a VGA (Video Graphics Array) analog connector (also known as a D-Sub connector) that has 15 pins in three rows. Typically blue in color.

What is A Computer?

<u>Ports –</u> Means of connecting peripheral devices to your computer.

- **Parallel Port** – Most often used to connect a printer to the computer. 25-pin connector. Long and skinny, often pink in color. Transmits data at 50-100 Kb/s.
- **USB Port –** Universal Serial Bus. Now used to connect almost all peripheral devices to the computer. USB 1.1 transmits data at 1.5 Mb/s at low speed, 12 Mb/s at full speed. USB 2.0 transmits data at 480 Mb/s.

What is A Computer?

<u>Ports –</u> Means of connecting peripheral devices to your computer.

- **PS/2 Port-** sometimes called a mouse port, was developed by IBM. It is used to connect a computer mouse or keyboard. Most computers come with two PS/2 ports.
- **Ethernet Port–** This port is used for networking and fast internet connections. Data moves through them at speeds of either 10 megabits or 100 megabits or 1 gigabit (1,000 megabits) depending on what speed the network card in the computer supports. Little monitor lights on these devices flicker when in use.

What is A Computer?

Power Supply – Gives your computer power by converting alternating current (AC) supplied by the wall connection to direct current (DC).

What is A Computer?

<u>**Expansion Cards**</u> – Used to add/improve functionality to the computer.

Sound Card – Used to input and output sound under program control. Sound cards provide better sound quality than the built in sound control provided with most computers.

Graphics Card – Used to convert the logical representation of an image to a signal that can be used as input for a monitor.

Network Card – Used to provide a computer connection over a network. Transmit data at 10/100/1000 Mb/s.

What is A Computer?

CD ROM – A device used to read CD-ROMs. If capable of writing to the CD-ROM, then they are usually referred to as a 'burner' or CD-RW.

DVD ROM – A device that is used to read DVDs/CDs. If capable of writing to the DVD, then it is often referred to as a DVD-burner or a DVD-RW.

Floppy Drive – A device that is used to read/write to floppy diskettes.

What is A Computer?

Fan – Keeps your computer cool. If the inside of your computer becomes too hot, then the computer can overheat and damage parts.

Heatsink – Used to disperse the heat that is produced inside the computer by the CPU and other parts by increasing surface area.

What is A Computer?

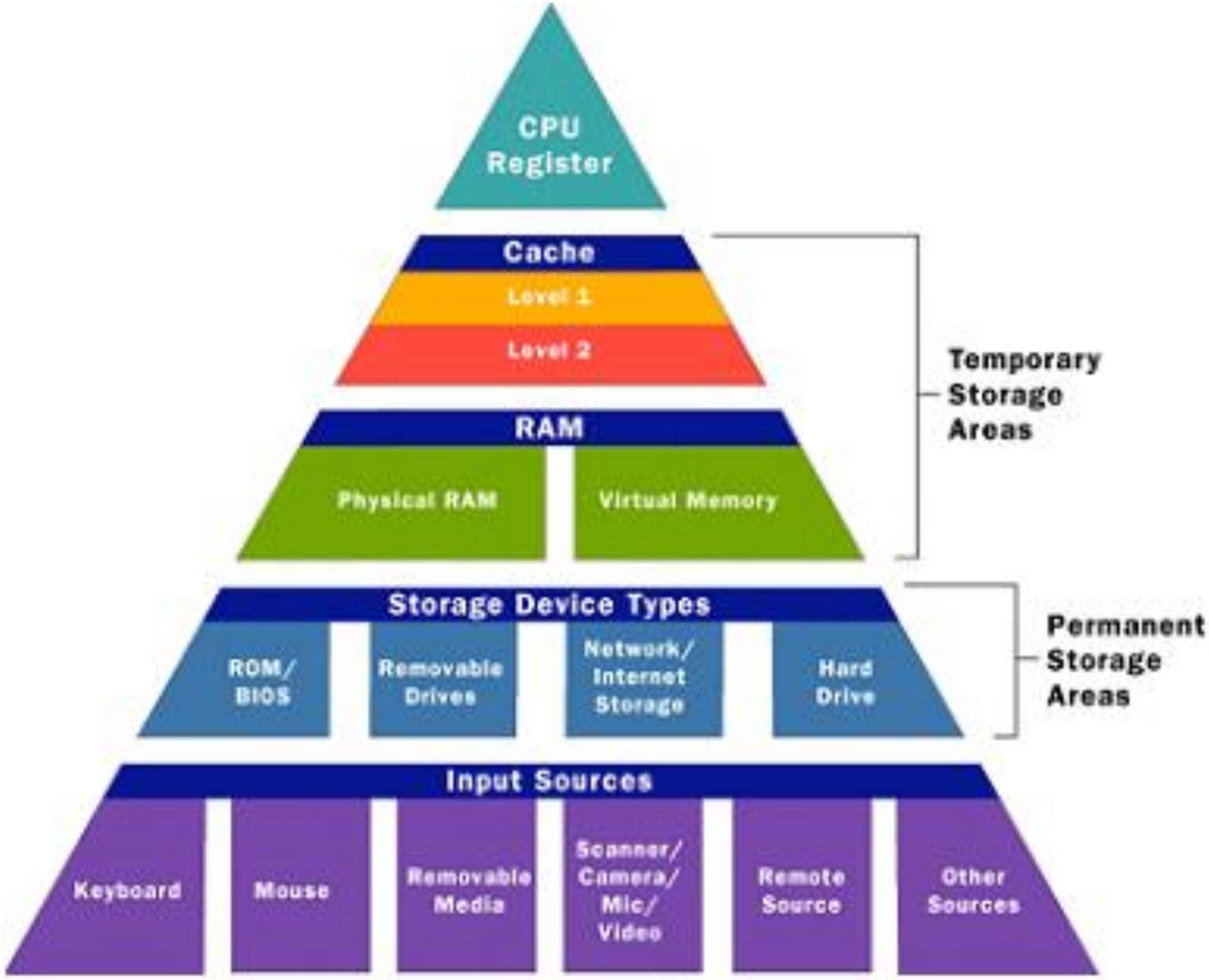
The little parts– Capacitors – store energy, Resistors – allows a current through, Transistors – a valve which allows currents to be turned on or off.

Case – (Tower if standing upright.) What your motherboard, CPU, etc is contained in.

Comprehension Questions

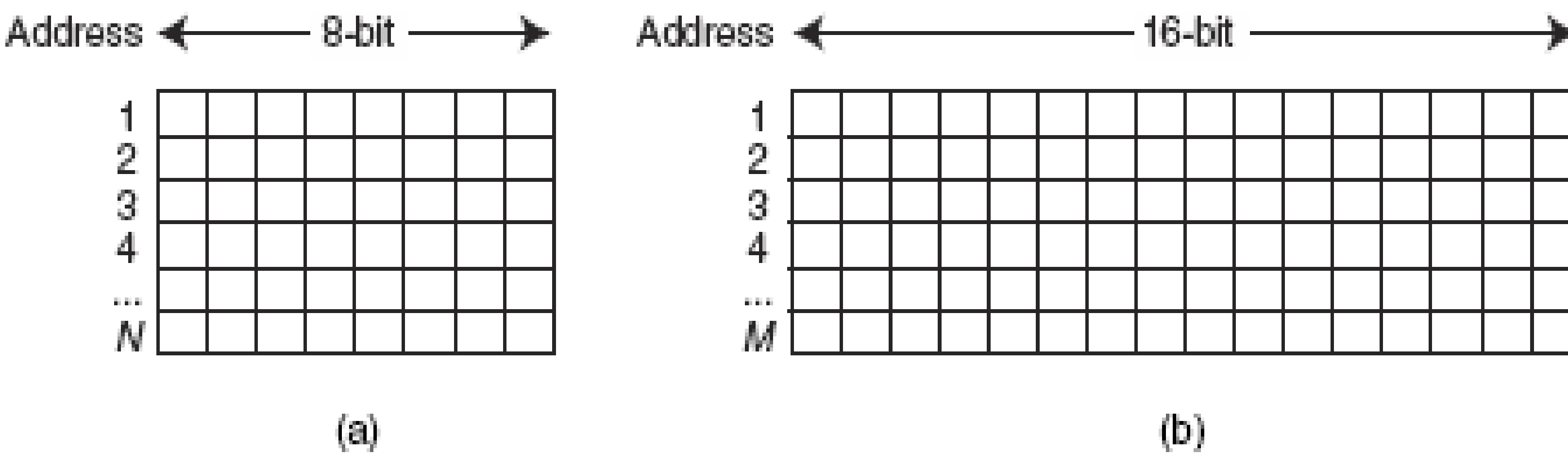
1. What are the 3 main components of a computer?
2. Name 3 input devices. Name 3 output devices.
3. What is the brain of the computer?
4. Explain the difference between memory and your hard drive.
5. What are the similarities and differences between RAM, ROM, and hard drives?
6. Describe each of the different ports.
8. What gives your computer power?

Memory


CPU Register
Cache
Level 1
Level 2
RAM
Physical RAM
Virtual Memory
Temporary Storage Areas
Storage Device Types
ROM/ BIOS
Removable Drives
Network/ Internet Storage
Hard Drive
Permanent Storage Areas
Input Sources
Keyboard
Mouse
Removable Media
Scanner/ Camera/ Mic/ Video
Remote Source
Other Sources

Memory Organization and Addressing

- **You can envision memory as a matrix of bits. Each row, implemented by a register, has a length typically equivalent to the word size of the machine. Each register (more commonly referred to as a *memory location*) has a unique address; memory addresses usually start at zero and progress upward.**



- **An address is almost always represented by an unsigned integer, 4-bits is a nibble, and 8- bits is a byte.**
- **Normally, memory is *byte-addressable* which means that each individual byte has a unique address. Some machines may have a word size that is larger than a single byte.**

- For example, a computer might handle 32-bit words (which means it can manipulate 32 bits at a time through various instructions), but still employ a byte-addressable architecture.

- In this situation, when a word uses multiple bytes, the byte with the lowest address determines the address of the entire word.

- It is also possible that a computer might be *word-addressable*, which means each word (not necessarily each byte) has its own address, but most current machines are byte-addressable (even though they have 32-bit or larger words).

- A memory address is typically stored in a single machine word

Internal Memory

- **<u>Semiconductor Memory</u>**
- It is all a random access memory.
- The main element of semiconductor memory is cell as shown in figure **<u>Cell properties</u>**
- They exhibit two states 0 or 1.
- Capable of being written into.
- Capable of being read from.

Semiconductor Memory

- RAM
  - Misnamed as all semiconductor memory is random access
  - Read/Write
  - Volatile
  - Temporary storage
  - Static or dynamic

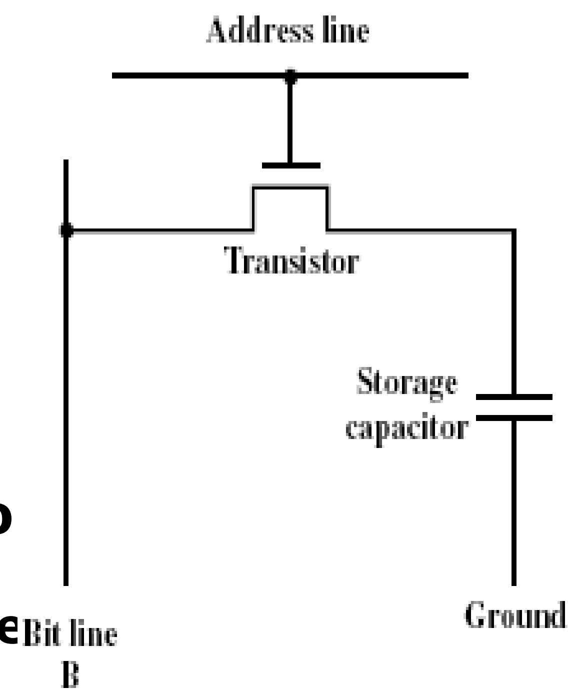
Dynamic RAM

- Bits stored as charge in capacitors
- Charges leak
- Need refreshing even when powered
- Simpler construction
- Smaller per bit
- Less expensive
- Need refresh circuits
- Slower
- Main memory
- Essentially analogue
  - Level of charge determines value

RAM (Random Access Memory)

DRAM

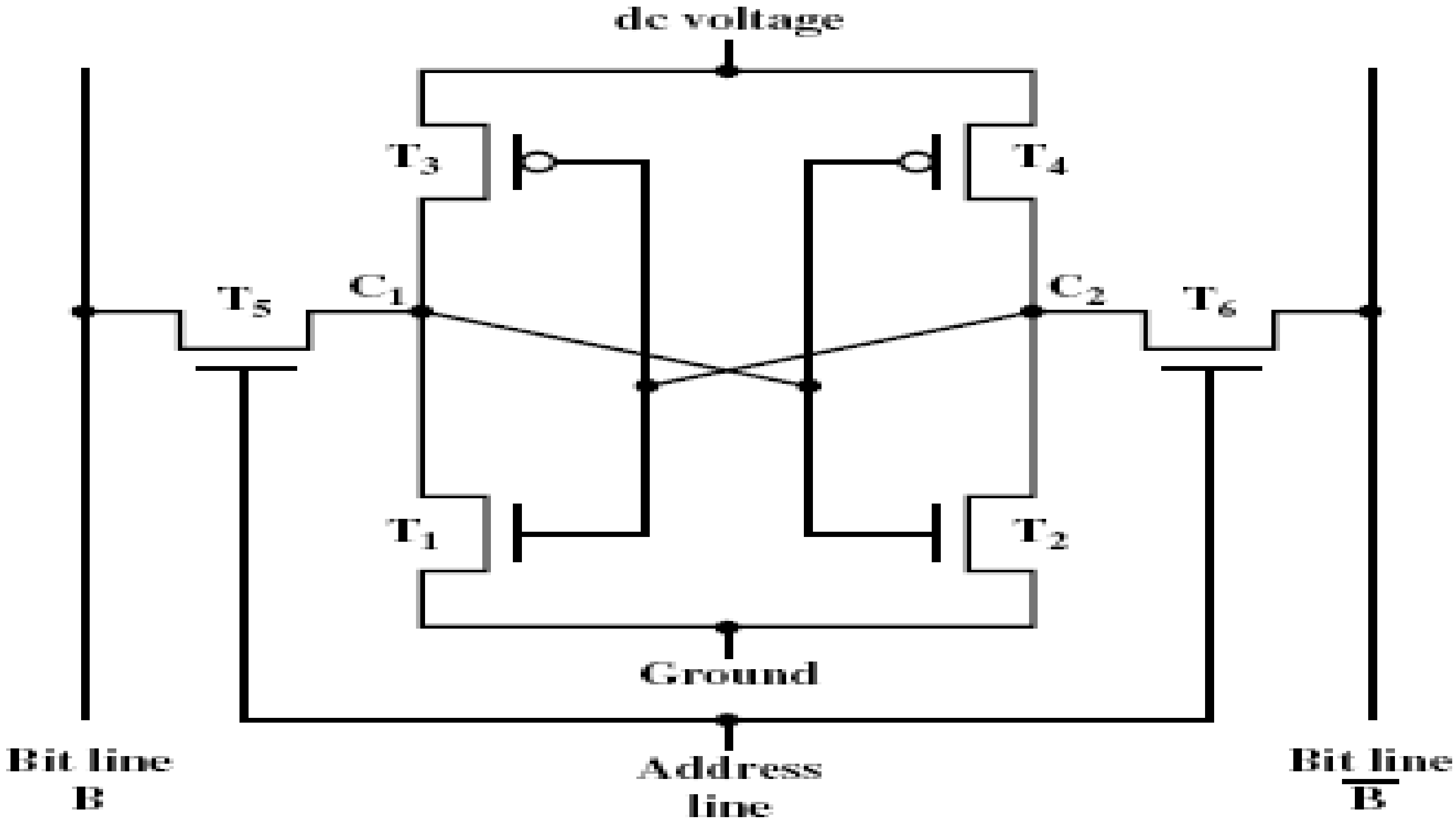
- **Build on the idea of charging a capacitor**
- **DRAM Operation**
- **Address line active when bit read or written**
  - **Transistor switch closed (current flows)**
- **Write**
  - **Voltage to bit line**
    - **High for 1 low for 0**
  - **Then signal address line**
    - **Transfers charge to capacitor**
- **Read**
  - **Address line selected**
    - **transistor turns on**
  - **Charge from capacitor fed via bit line to sense amplifier**
  - **Compares with reference value to determine 0 or 1**



- **Capacitor charge must be restored**

Static RAM Operation

- Transistor arrangement gives stable logic state
- State 1
  - C_1 high, C_2 low
  - T_1 T_4 off, T_2 T_3 on
- State 0
  - C_2 high, C_1 low
  - T_2 T_3 off, T_1 T_4 on
- Address line transistors T_5 T_6 is switch
- Write – apply value to B & compliment to B
- Read – value is on line B



Address line transistors T5 T6 is switch
Write – apply value to B & compliment to B
Read – value is on line.

Static RAM

- Bits stored as on/off switches
- No charges to leak
- No refreshing needed when powered
- More complex construction
- Larger per bit
- More expensive
- Does not need refresh circuits
- Faster
- Cache
- Digital
  - Uses flip-flops

SRAM v DRAM

- Both volatile
  - Power needed to preserve data
- Dynamic cell
  - Simpler to build, smaller
  - More dense
  - Less expensive
  - Needs refresh
  - Larger memory units
- Static
  - Faster
  - Cache

Comparison between DRAM and SRAM

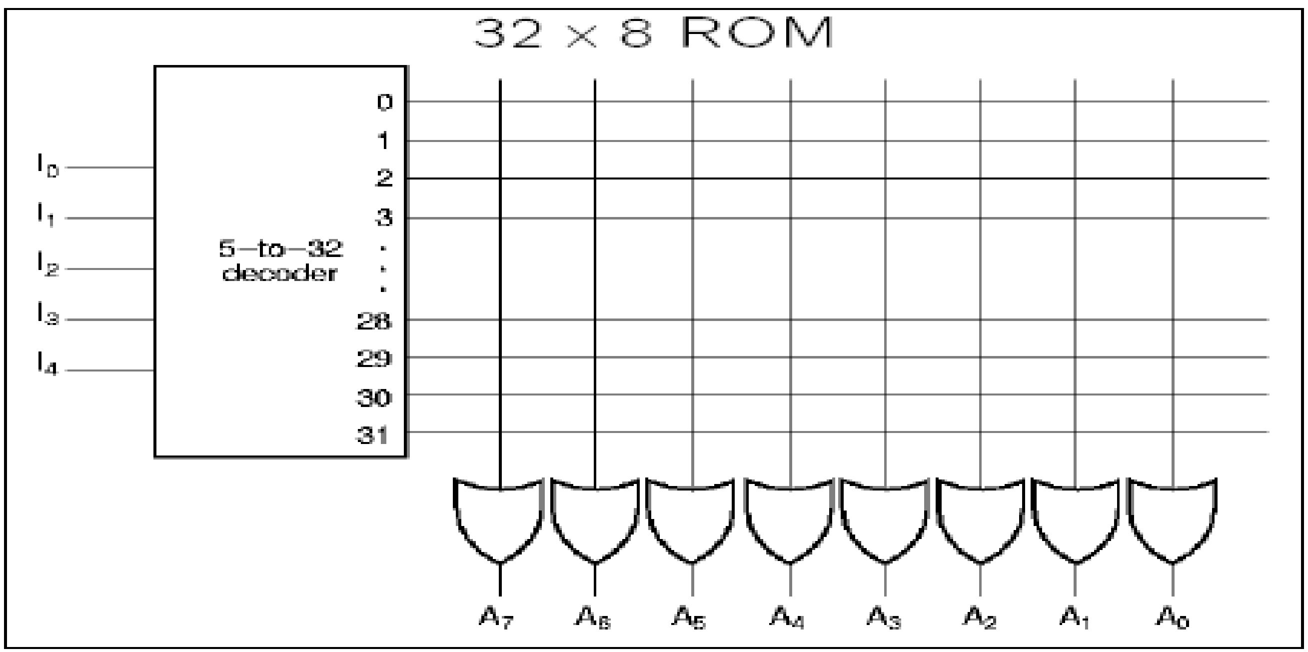
DRAM	SRAM
Simpler to build, smaller	Larger than DRAM
More dense	Faster
Less expensive	More expensive
Needs refresh	Doesn't need refresh
Larger memory units (Main Memory)	Smaller memory units (Cache)

ROM (Read Only Memory)

- The programming here is by hardware
- Non-volatile

Read-Only Memory

k inputs (address) $\Rightarrow$ $\boxed{2^k \times n \text{ ROM}}$ $\Rightarrow$ n outputs (data)



Types of ROM

PROM (Programmable Read Only Memory)

- Programmed for once using special equipment, and never erased.

EPROM (Erasable Programmable Read Only Memory)

- Programmable and erased using ultra violet.
- Erased as a whole.

EEPROM (Electrically Erasable Programmable Read Only Memory)

- Programmable and erased using electrical shots.
- Erased byte by byte
- Can erase a part without another

Flash memory

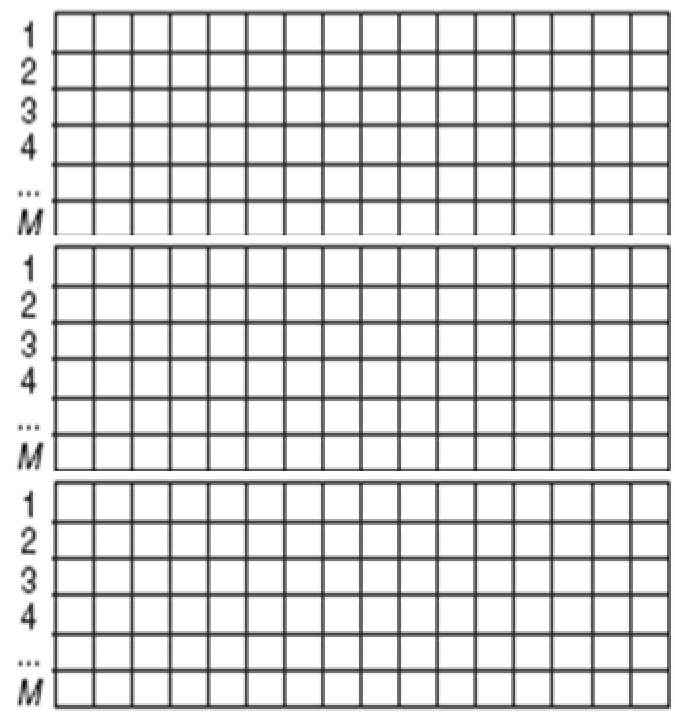
- Programmable and erased using electrical shots.
- Erased block by block
- Erased so quick (in a flash).

Chip organization

- The question is how to put the memory on a chip…… ??
- If we have a memory of 2^n location with 32-bits each, so, it can be represented by one of two means:-

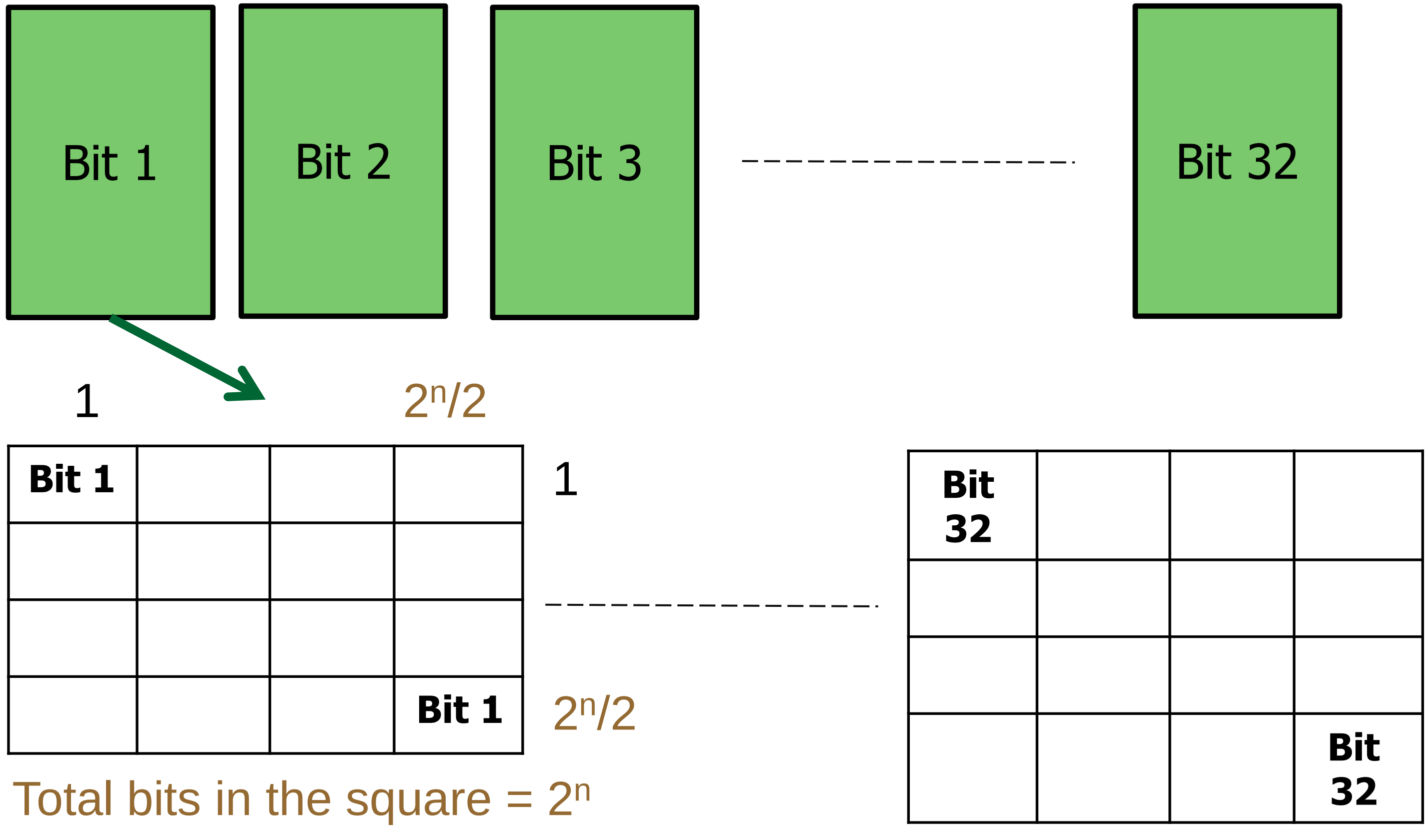
1- Put it in the form of a matrix with 2^n rows and 32 columns.

→But this architecture is not suitable because the length is much more than its width.

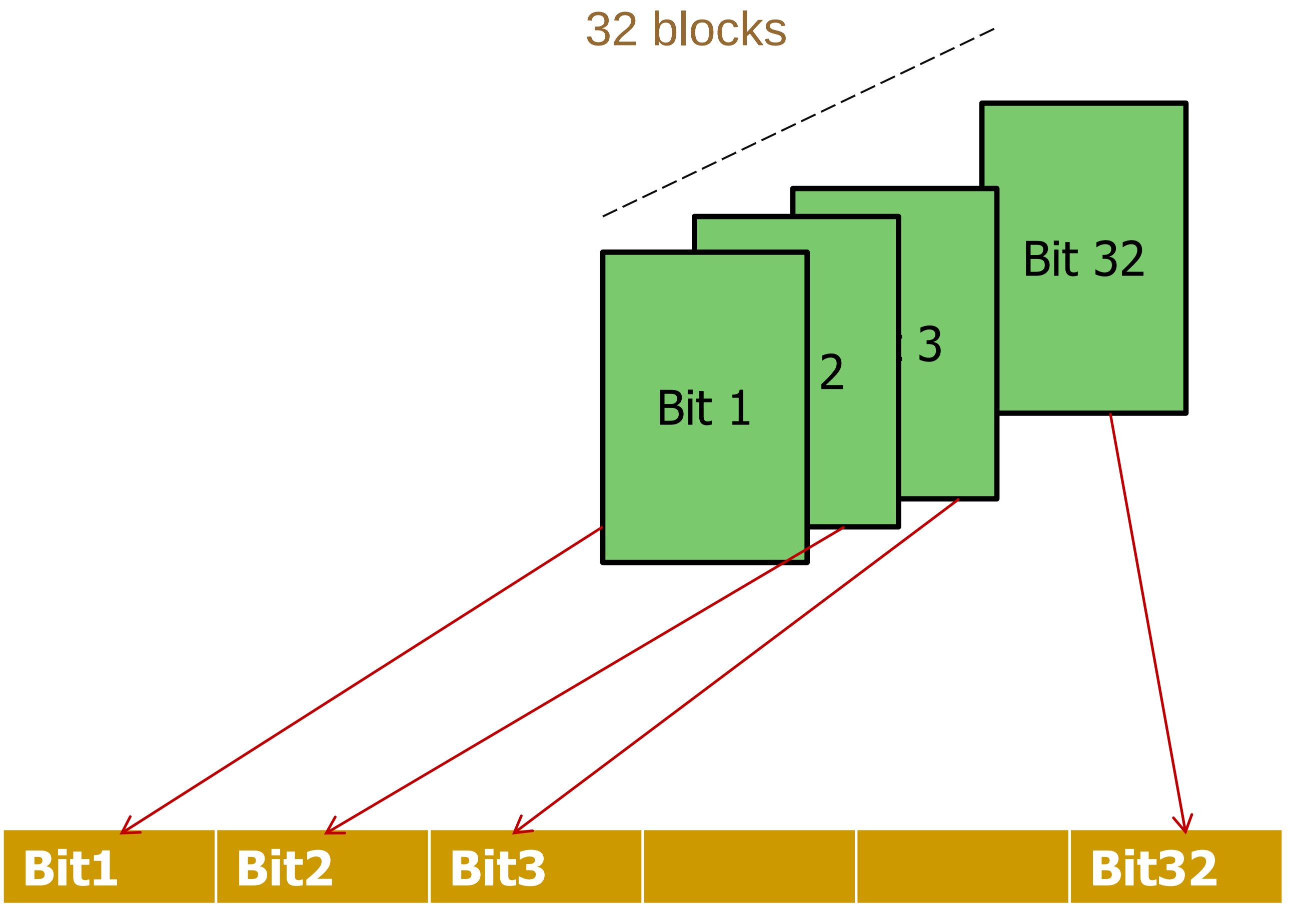


Use a group of blocks each will represent a bit, then we need 32 blocks.

- Each block contains 2^n bits represented as square matrix: $2^n/2$ rows and $2^n/2$ columns.
- The size will be smaller than the first solution.



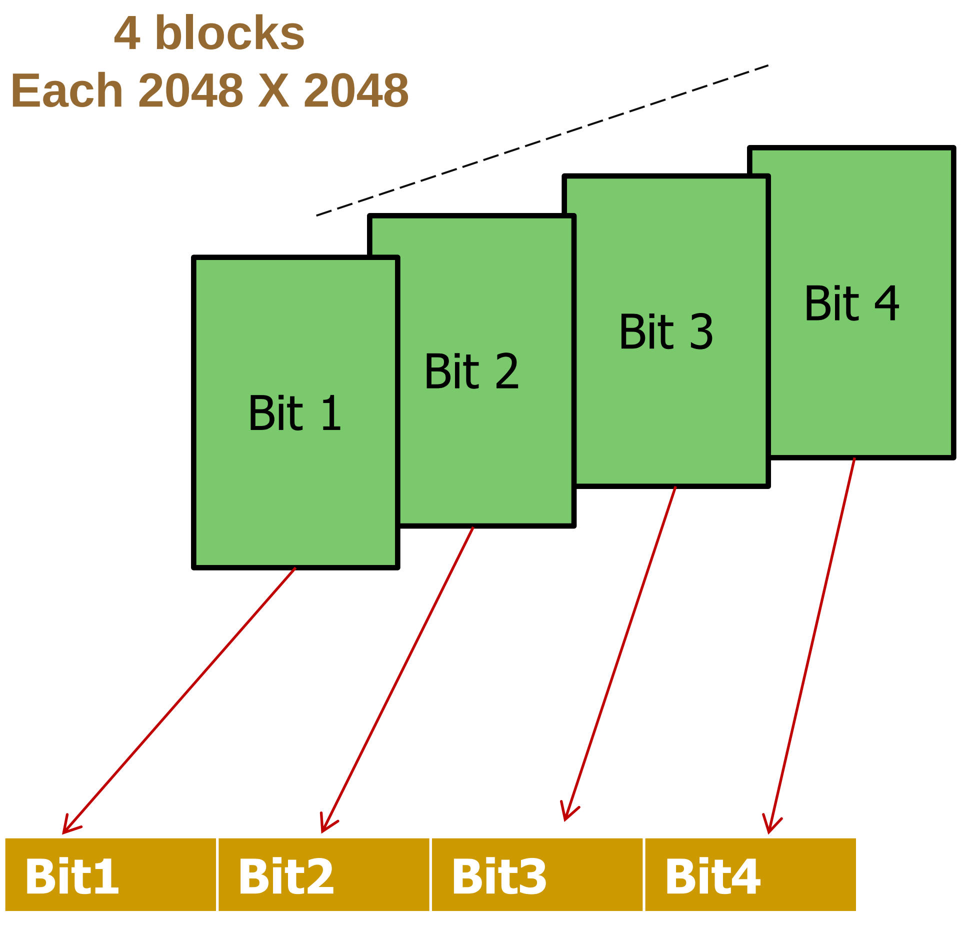
Total bits in the square = 2^n



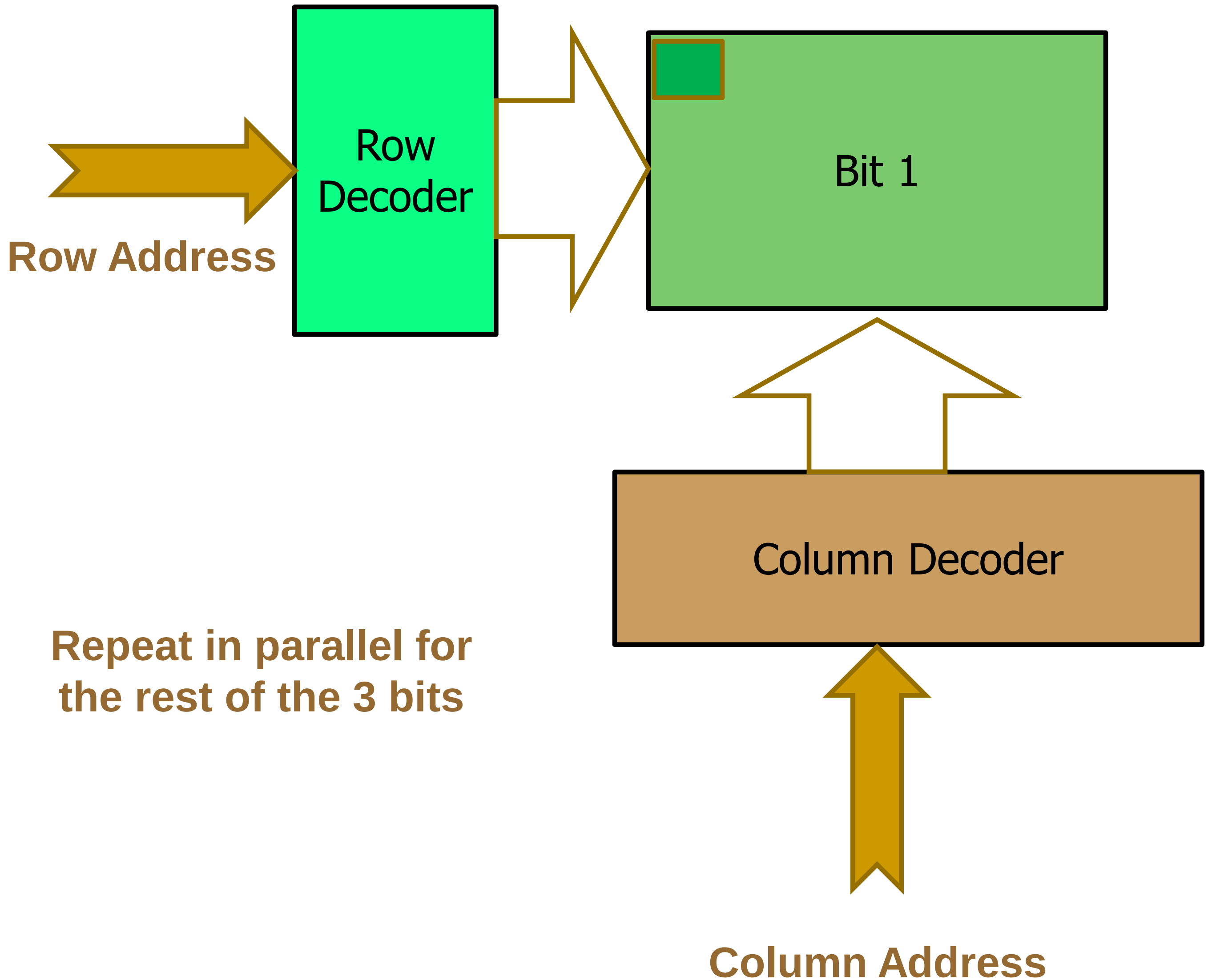
Bits are collected each from each block

Example:

- A 16-Mbit DRAM (4 Mbits each 4 bits word) can be organized so that 4-bits are read or written at a time
- So logically the memory array is organized as 4 square arrays of 2048 (2 Kbits) by 2048 (2 Kbits) elements.
- The elements of the array are connected by both horizontal (row) and vertical (columns) lines.
- We need 22 bits address (4 Mbits = 2^{22} bits)
- In this configuration, there are only 11-address lines(A0-A10), half the number you would expects for 2048 by 2048 array.
- We will enter it 11 bits by 11 bits.

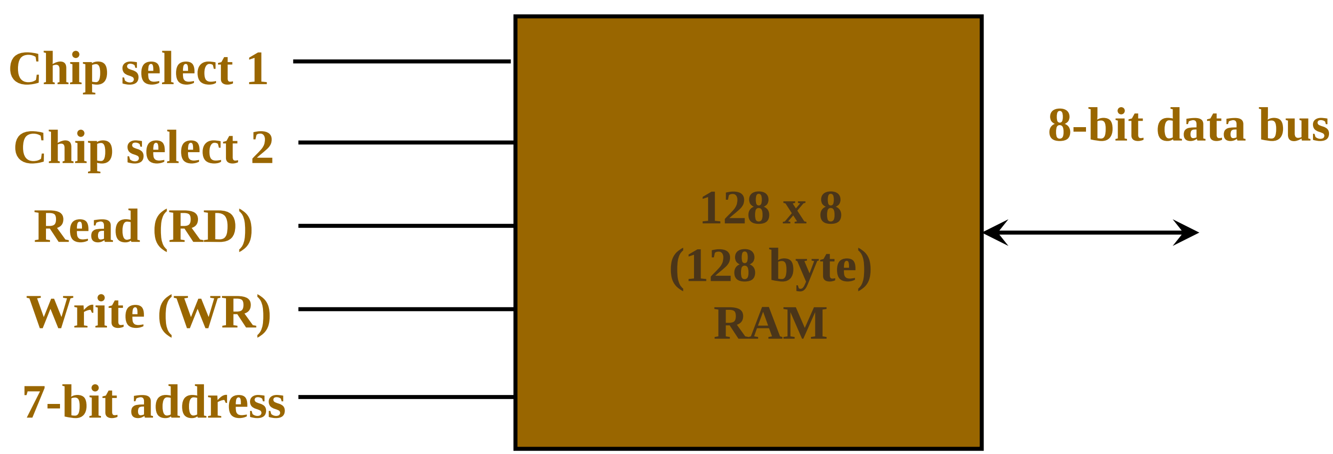


The four Bits are collected each from its corresponding block


Row Decoder
Bit 1
Row Address
Column Decoder
Repeat in parallel for the rest of the 3 bits
Column Address

RAM Chips

- A RAM chip is faster when communicating with CPU than dealing with Auxiliary devices directly.
- If the memory needed for the computer is larger than the capacity of one RAM chip, we combine a number of chips to form required memory size.
- If we have many RAM chips, we must choose to access one of them
- We use control inputs to select the chip only when needed.



$128 = 2^7$ **Address = 7 bits**

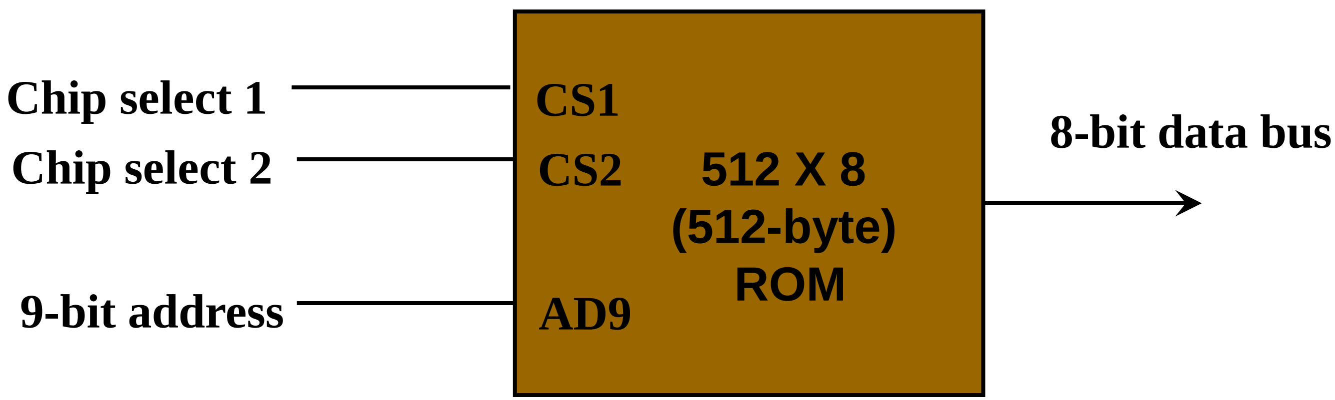
8 = 1 byte of data

RAM chip function table

CS1	CS2	RD	WR	Memory function	State of data bus
0	0	X	X	Inhibit	High-impedance
0	1	X	X	Inhibit	High-impedance
1	0	0	0	Inhibit	High-impedance
1	0	0	1	Write	Input data to RAM
1	0	1	X	Read	Output data from RAM
1	1	X	X	Inhibit	High-impedance

ROM Chips

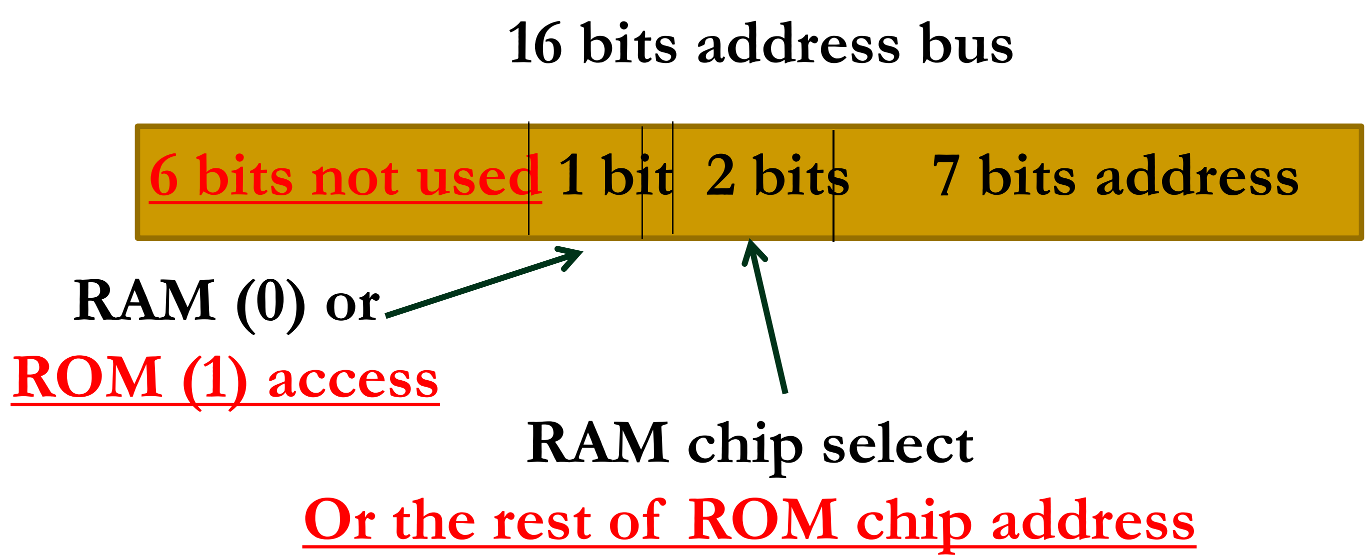
- A ROM chip is organized externally in a similar manner. ROM can only read, the data bus can only be in an output mode.
- For the same–size chip, it is possible to have more bits of ROM than of RAM, because the internal binary cells in ROM occupy less space than in RAM.
- For this reason, the diagram specifies a 512-byte ROM, while the RAM has only 128 bytes.
- address lines = 9 bits ($512 = 2^9$)
- The two chip select inputs must be CS1=1 and CS2= 0 for the unit to operate.
- Otherwise, the data bus is in a high-impedance state.

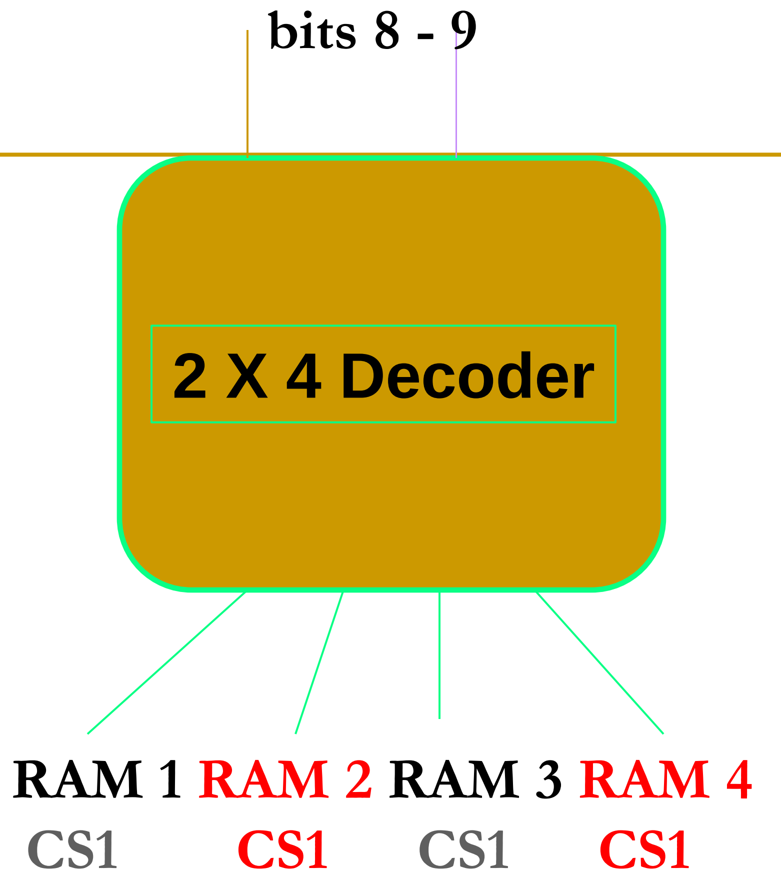


Practical Example : How CPU deal with RAM and ROM

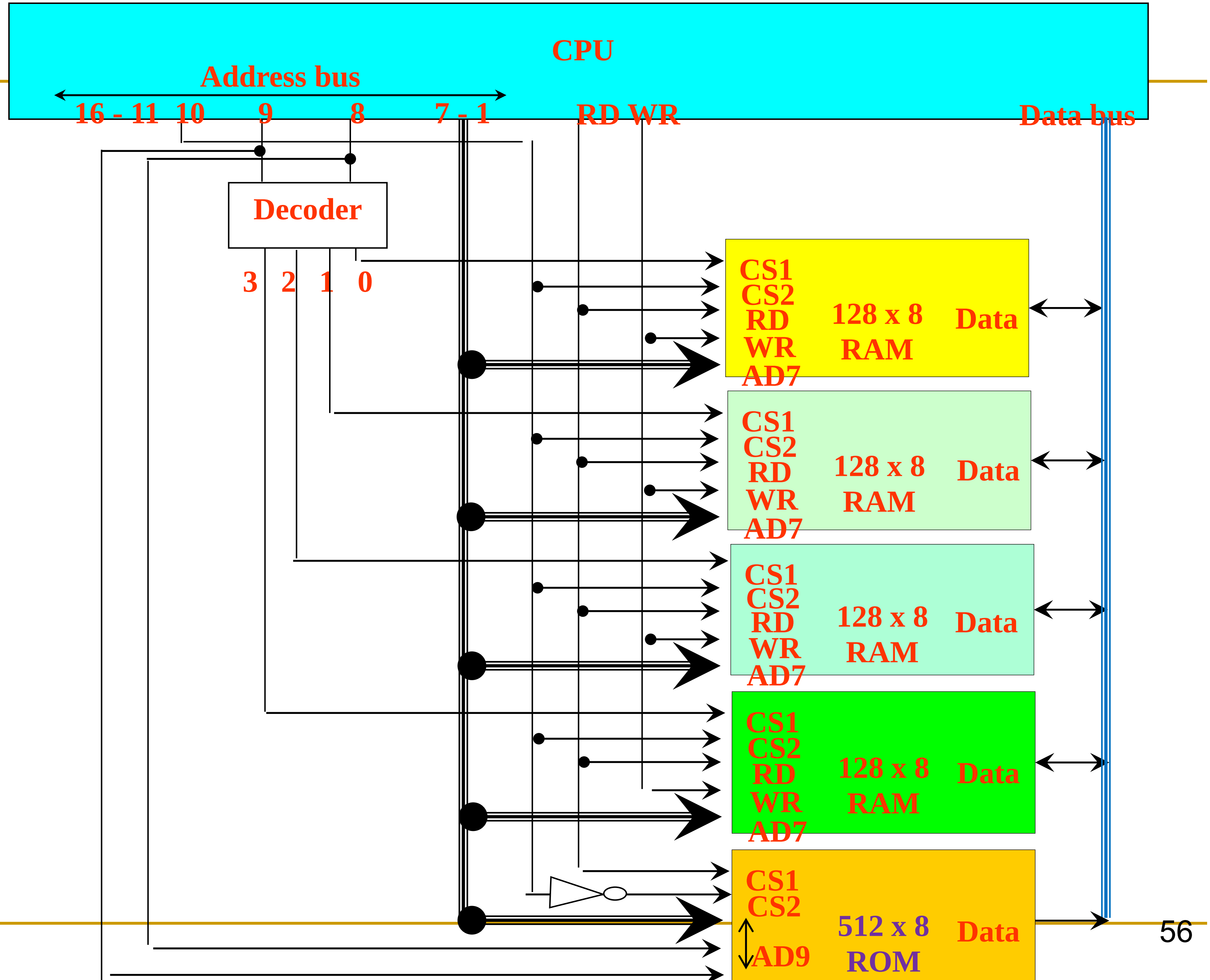
Assume that a computer system has:

- 512 bytes of RAM (we use 4 blocks of RAM of the same type 128X8)
- 512 bytes of ROM (single block).





We use a decoder to select one of the 4 RAM chips using bits 8 – 9 by connecting the output to CS1 of the RAM Chip


CPU
Address bus
16 - 11
10
9
8
7 - 1
RD WR
Data bus
Decoder
3
2
1
0
CS1
CS2
RD
WR
AD7
128 x 8
RAM
Data
CS1
CS2
RD
WR
AD7
128 x 8
RAM
Data
CS1
CS2
RD
WR
AD7
128 x 8
RAM
Data
CS1
CS2
RD
WR
AD7
128 x 8
RAM
Data
CS1
CS2
AD9
512 x 8
ROM
Data

Component	Hexadecimal address	Address bus									
		10	9	8	7	6	5	4	3	2	1
RAM 1	0000-007F	0	0	0	X	X	X	X	X	X	X
RAM 2	0080-00FF	0	0	1	X	X	X	X	X	X	X
RAM 3	0100-017F	0	1	0	X	X	X	X	X	X	X
RAM 4	0180-01FF	0	1	1	X	X	X	X	X	X	X
ROM	0200-03FF	1	X	X	X	X	X	X	X	X	X

$(0000\ 0000\ 0000\ 0000)_2 = (0000)_{16}$

$(0000\ 0000\ 0111\ 1111)_2 = (007F)_{16}$

$(0000\ 0000\ 1000\ 0000)_2 = (0080)_{16}$

$(0000\ 0000\ 1111\ 1111)_2 = (00FF)_{16}$

$(0000\ 0001\ 0000\ 0000)_2 = (0100)_{16}$

$(0000\ 0001\ 0111\ 1111)_2 = (017F)_{16}$

Thanks for your attention